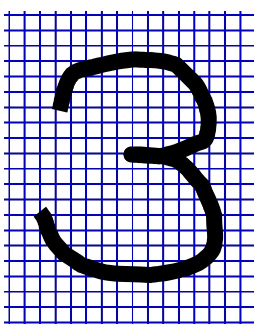
サポートベクトル分類（８章）

杉山将・本多淳也
sugi@k.u-tokyo.ac.jp, jhonda@k.u-tokyo.ac.jp
http://www.ms.k.u-tokyo.ac.jp

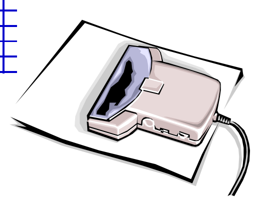
パターン認識

- 入力パターンをカテゴリに割り当てる**識別関数**を構成する問題
  - 問題に合わせて人間が識別関数を設計
  - **データから自動的に識別関数を学習**

パターン



$x \in \mathbf{R}^d$



→ 識別関数 $y = f(x)$ →

カテゴリ

3

$y \in \{1, 2, \ldots, c\}$

統計的パターン認識

- **訓練標本**: 属するカテゴリが既知のパターン

$$\{(x_i, y_i)\}_{i=1}^{n}$$

$$x_i \in \mathbf{R}^d$$

$$y_i \in \{1,2,\ldots,c\}$$

- **統計的パターン認識**: 訓練標本の統計的な性質を利用して識別関数を学習する

- **仮定**:

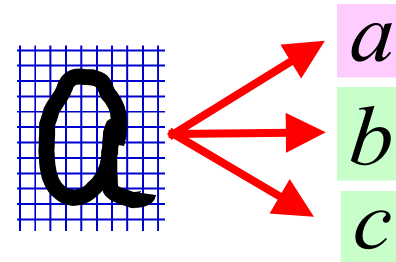
$$(x_i, y_i) \overset{\text{i.i.d.}}{\sim} p(x, y)$$

i.i.d. (independent and identically distributed)

独立に同一の分布に従う

理想的なパターン分類法

- 事後確率 $p(y \mid x)$: 与えられたパターン x がクラス y に属する確率

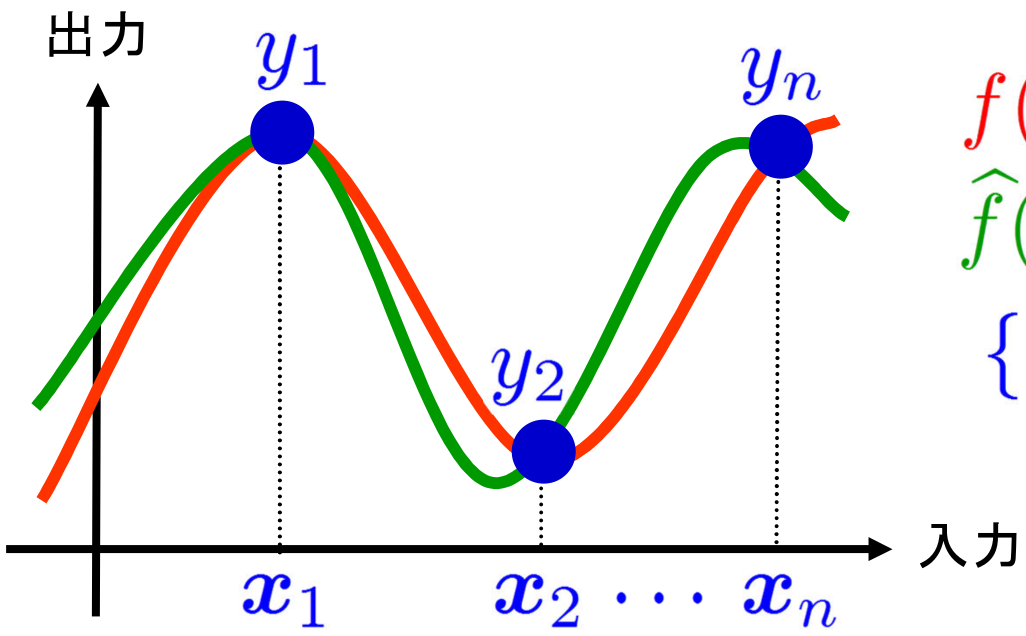


- 事後確率を最大にするカテゴリにパターンを分類すれば，パターンの誤識別率が最小になる．

$$f(x) = \arg\max_{y} \; p(y \mid x)$$

- 実際には事後確率は未知なので，訓練標本から推定しなければならない．

識別モデルの学習 ＝ 関数近似



$f(\boldsymbol{x})$:学習したい真の関数

$\widehat{f}(\boldsymbol{x})$:学習結果の関数

$\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$:訓練標本

$$y_i = f(\boldsymbol{x}_i)\ (+\mathrm{noise})$$

訓練標本から真の関数にできるだけ近い関数を求める

パターン認識では $y \in \{1, 2, \ldots, c\}$ であるが，
上記の図は $y \in \mathbb{R}$（回帰）に対応している．

パラメータに関する線形モデル

■ 線形モデル: $$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{j=1}^{b} \theta_j \phi_j(\boldsymbol{x})$$

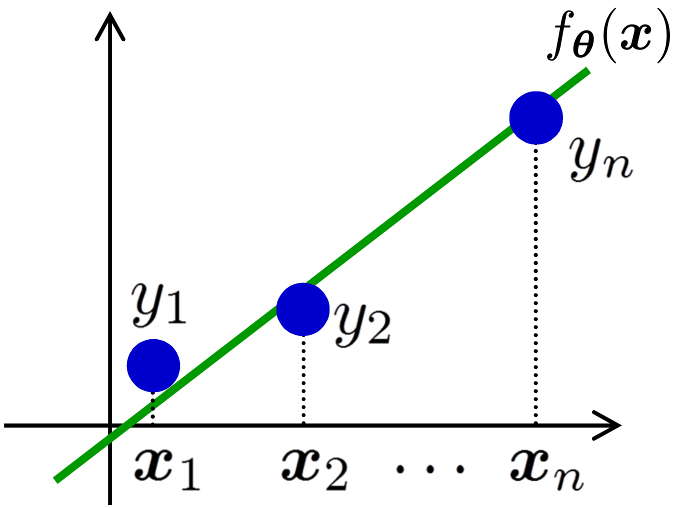
$\{\phi_j(\boldsymbol{x})\}_{j=1}^{b}$: 基底関数

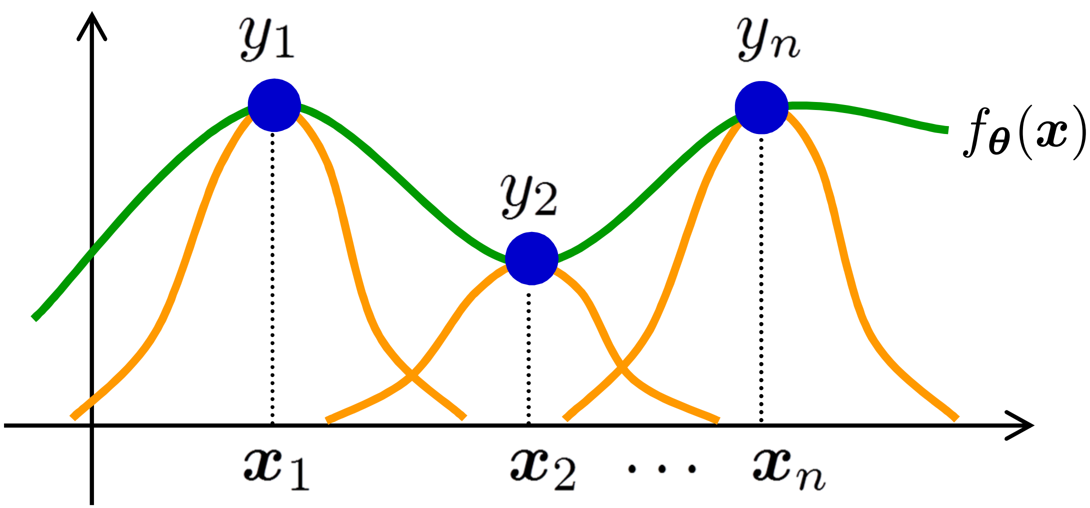
■ カーネルモデル:

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{j=1}^{n} \theta_j K(\boldsymbol{x}, \boldsymbol{x}_j)$$

ガウスカーネル

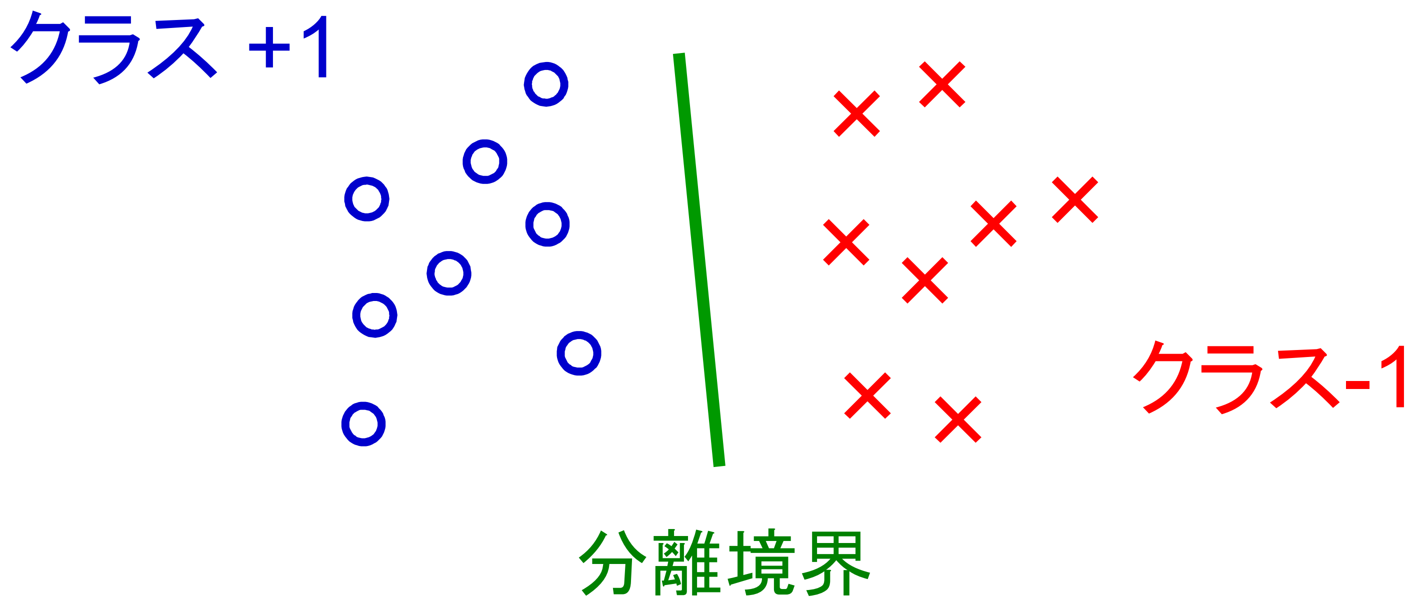
$$K(\boldsymbol{x}, \boldsymbol{c}) = \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{c}\|^2}{2h^2}\right)$$





2クラスの分類問題

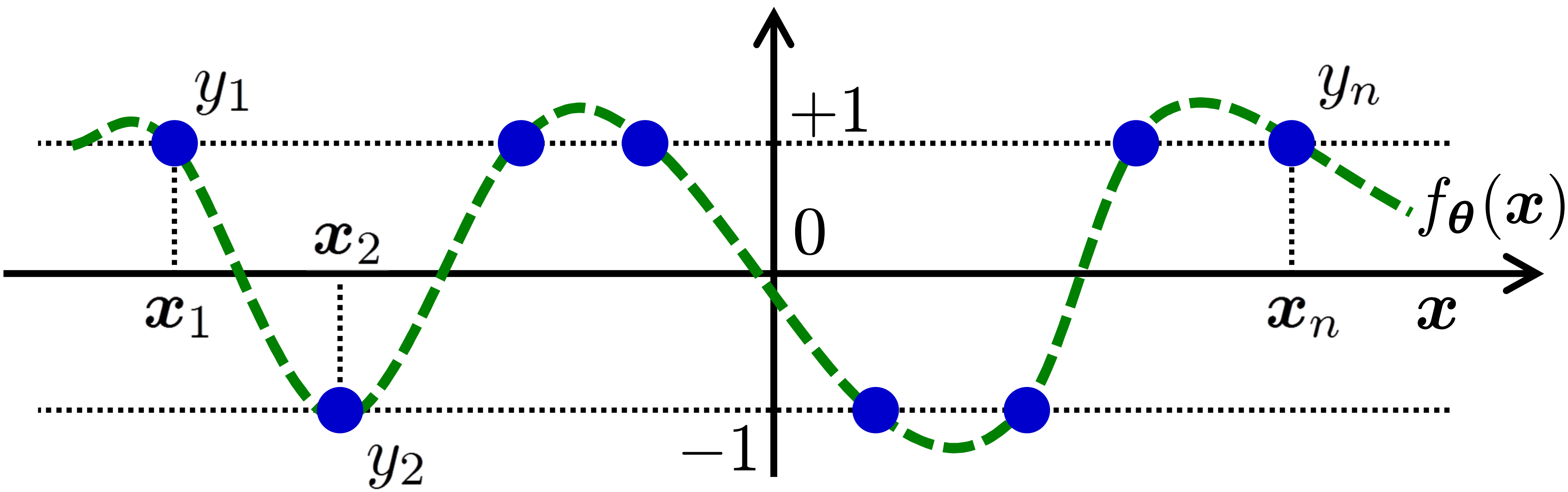
- ラベル付き訓練データ: $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$
  - 入力 $\boldsymbol{x}$ は d 次元の実ベクトル $\boldsymbol{x} \in \mathbb{R}^d$
  - 出力 y は2値のクラスラベル $y \in \{+1, -1\}$



- クラス間の分離境界を求めたい

2クラスの分類問題

- 2クラス分類問題は2値関数の近似問題と等価：



- 回帰学習法が分類にも使える！

回帰学習による分類

- パラメータを正則化最小二乗回帰で学習

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \left[\frac{1}{2} \sum_{i=1}^{n} \Big(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) - y_i \Big)^2 + \frac{\lambda}{2} \|\boldsymbol{\theta}\|^2 \right]$$

$\lambda\ (\geq 0)$: 正則化パラメータ

- テストパターンの分類:

$$\widehat{y} = \operatorname{sign}\left(f_{\widehat{\boldsymbol{\theta}}}(\boldsymbol{x})\right) = \begin{cases} +1 & (f_{\widehat{\boldsymbol{\theta}}}(\boldsymbol{x}) > 0) \\ 0 & (f_{\widehat{\boldsymbol{\theta}}}(\boldsymbol{x}) = 0) \\ -1 & (f_{\widehat{\boldsymbol{\theta}}}(\boldsymbol{x}) < 0) \end{cases}$$

0/1-損失関数とマージン

- 分類問題では，学習した関数の符号だけが必要

$$\widehat{y} = \mathrm{sign}\left(f_{\widehat{\boldsymbol{\theta}}}(\boldsymbol{x})\right)$$

- ℓ_2-損失でなく，0/1-損失の方が自然

$$J_{0/1}(\boldsymbol{\theta}) = \frac{1}{2}\sum_{i=1}^{n}\left(1 - \mathrm{sign}\left(m_i\right)\right)$$

$$m_i = f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)y_i$$

マージン

- $\mathrm{sign}(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)) = \mathrm{sign}(y_i)$ ⇒ $\mathrm{sign}(m_i) = 1$
- $\mathrm{sign}(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)) \neq \mathrm{sign}(y_i)$ ⇒ $\mathrm{sign}(m_i) = -1$

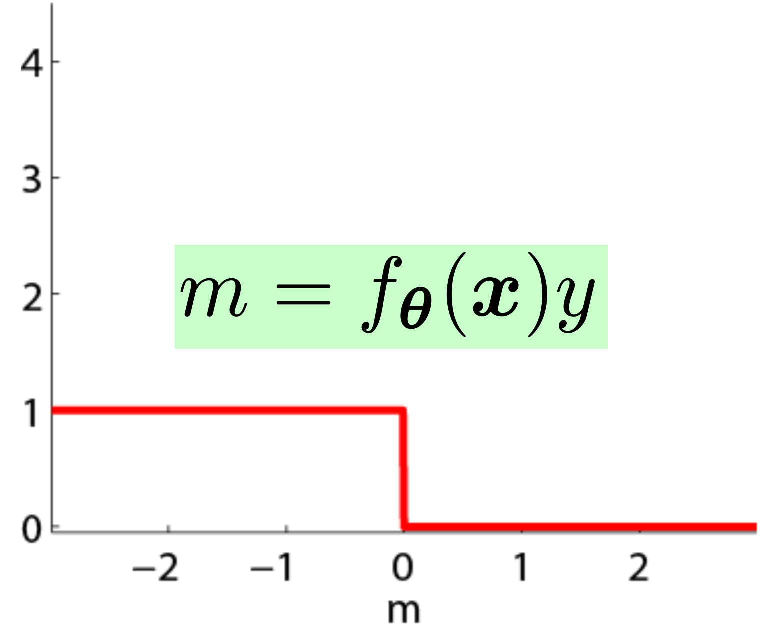
- $J_{0/1}(\boldsymbol{\theta})$ は誤分類標本数に相当．

0/1-損失関数とマージン

$$J_{0/1}(\boldsymbol{\theta}) = \frac{1}{2}\sum_{i=1}^{n}\left(1 - \mathrm{sign}\left(m_i\right)\right)$$

- 0/1-損失は誤分類標本数に対応
  - マージンが正なら誤差0
  - マージンが負なら誤差1
- 分類の損失としては理想的
  - しかし傾きを持たない離散的な関数
- 0/1-損失の最小化はNP困難
  - 現実的な時間では不可能

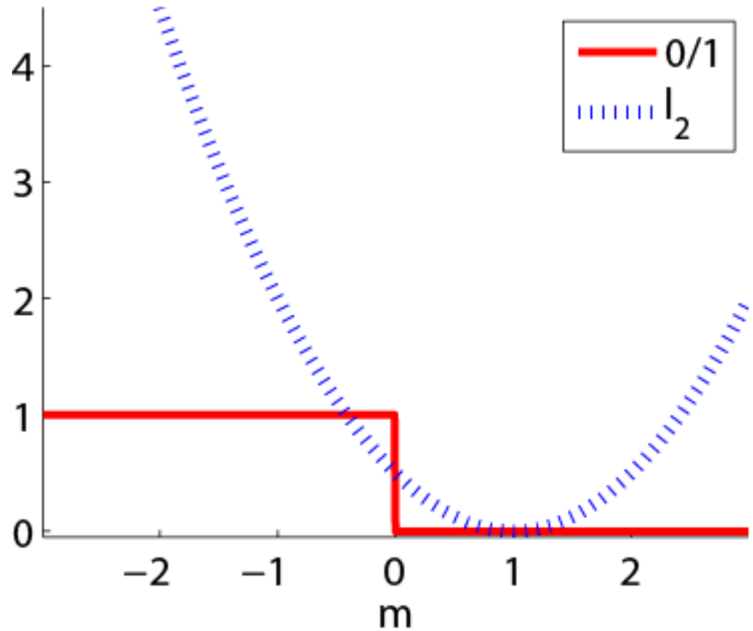
$$m = f_{\boldsymbol{\theta}}(\boldsymbol{x})y$$



ℓ_2-損失とマージン

$$\frac{1}{2}\Big(f_{\boldsymbol{\theta}}(\boldsymbol{x}) - y\Big)^2 = \frac{1}{2}(1-m)^2$$

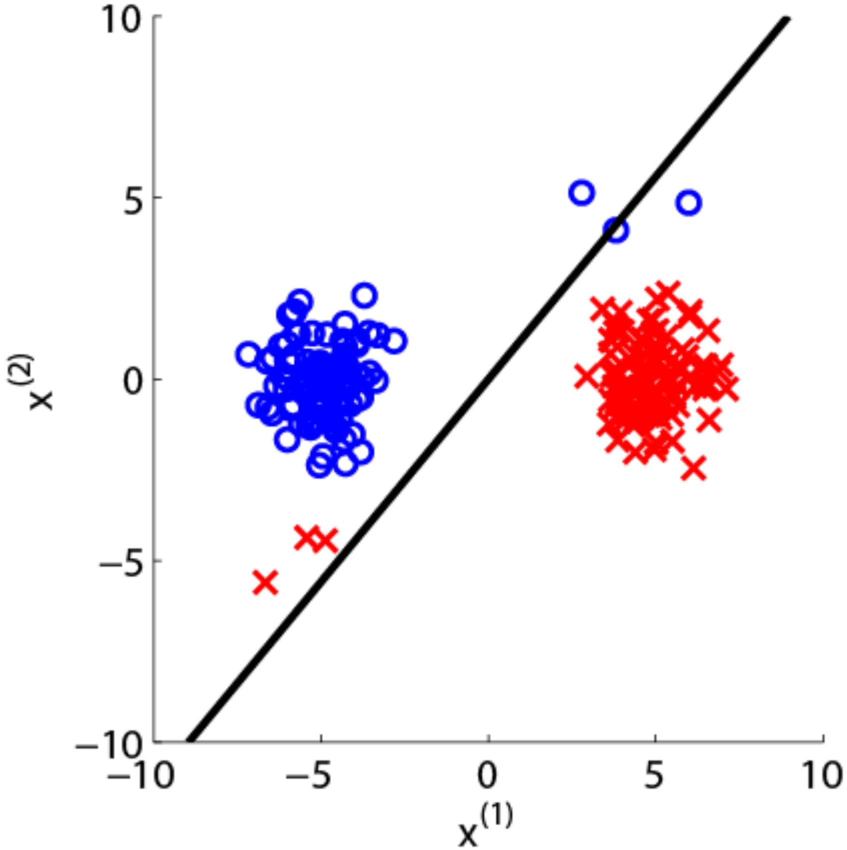
$$m = f_{\boldsymbol{\theta}}(\boldsymbol{x})y$$



- ℓ_2-損失関数は連続関数で扱いやすい
- 負のマージンを正(＋1)にしようとする
- 正の大きいマージンを＋1に減らそうとする

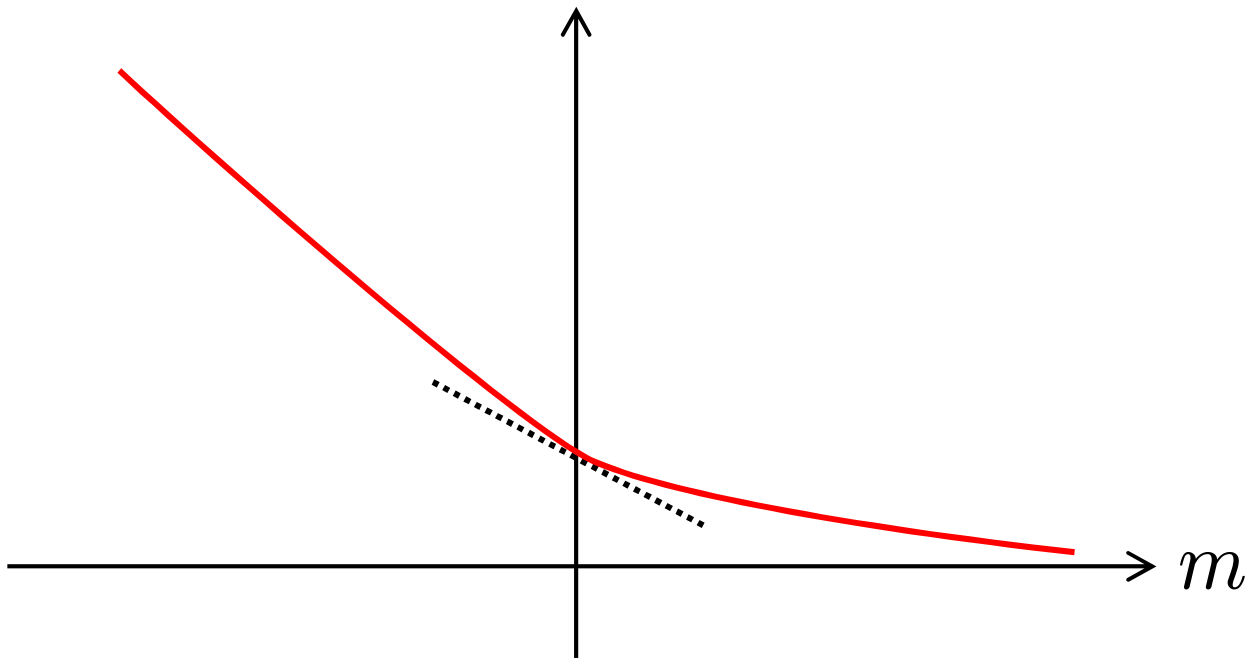
ℓ_2-損失の問題点

- 正の大きいマージンを＋1に減らそうとすることにより，下記のデータを正しく分離できない



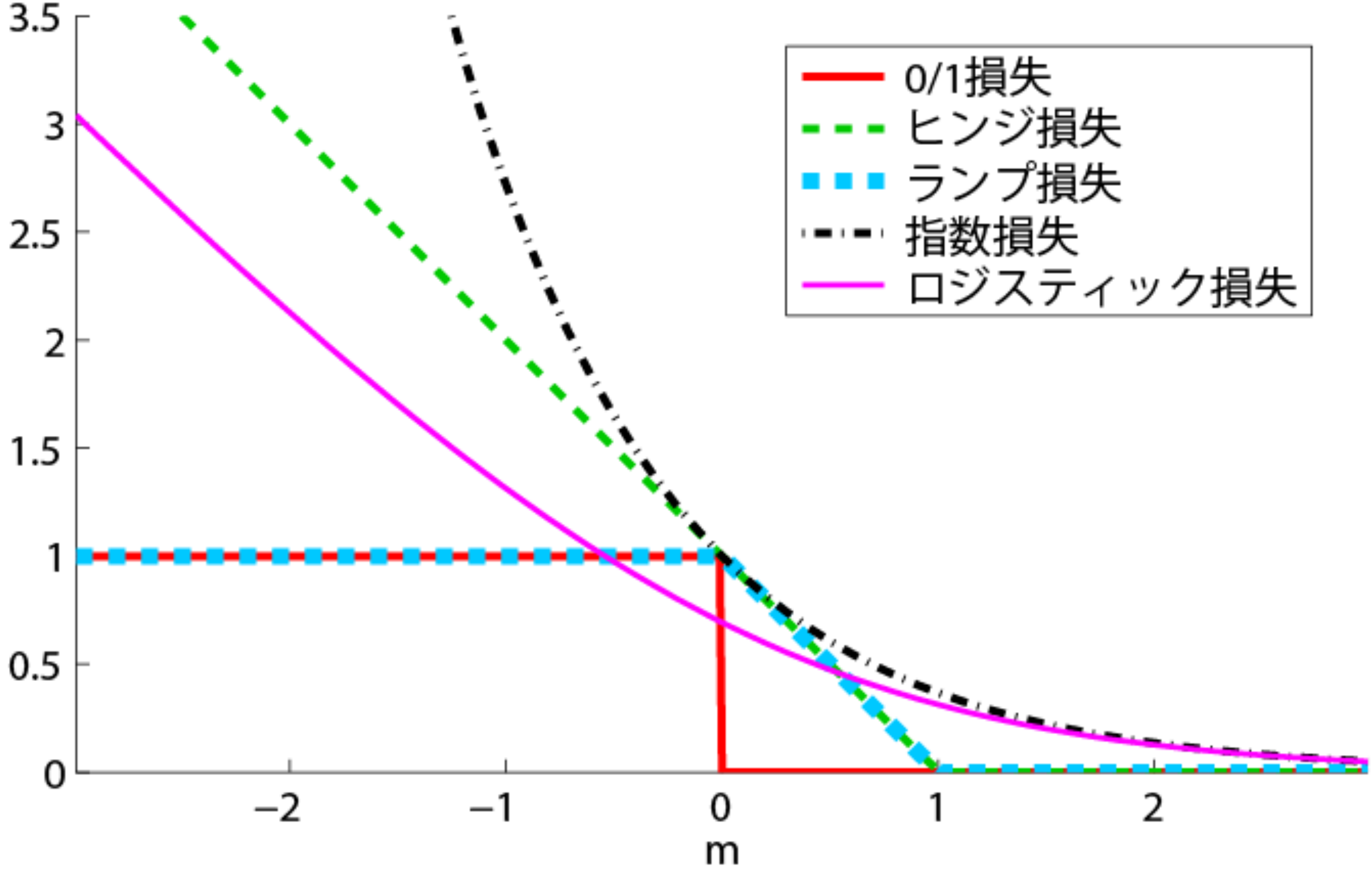
代理損失

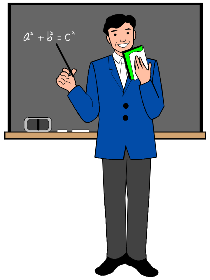
- 0/1-損失の代理として使う損失は, 単調非増加で $m = 0$ での傾きが負のものがよい
  - 負のマージンを正にしようとする
  - 正のマージンは減らさない



代理損失

- 機械学習では，様々な代理損失が用いられる





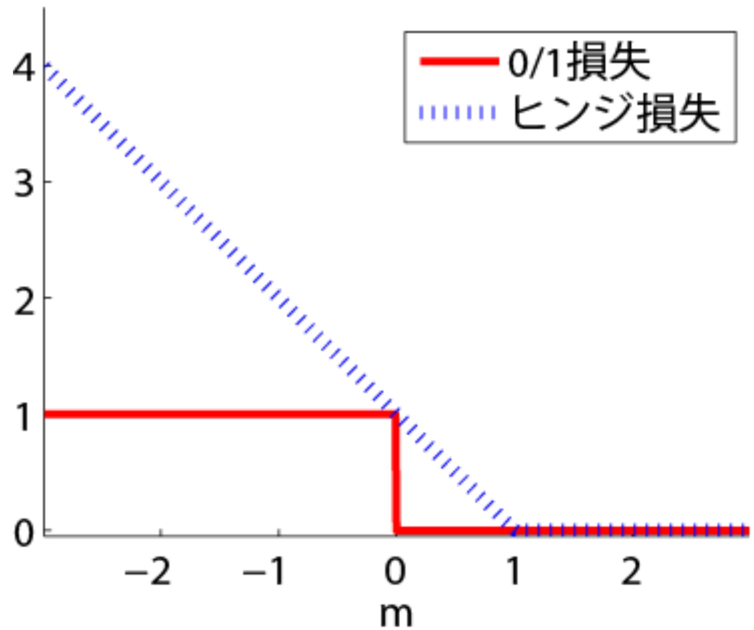
講義の流れ

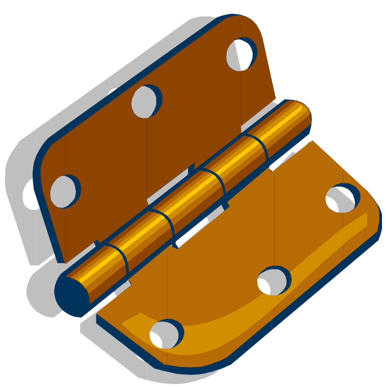
1. ヒンジ損失（8章）
2. サポートベクトル分類の元々の導出（8章）
3. サポートベクトル分類の解の性質（8章）

ヒンジ損失

- 0/1-損失の代理としてヒンジ損失を用いる

$$\sum_{i=1}^{n} \max(0, 1 - m_i) \qquad m_i = f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) y_i$$





サポートベクトル分類

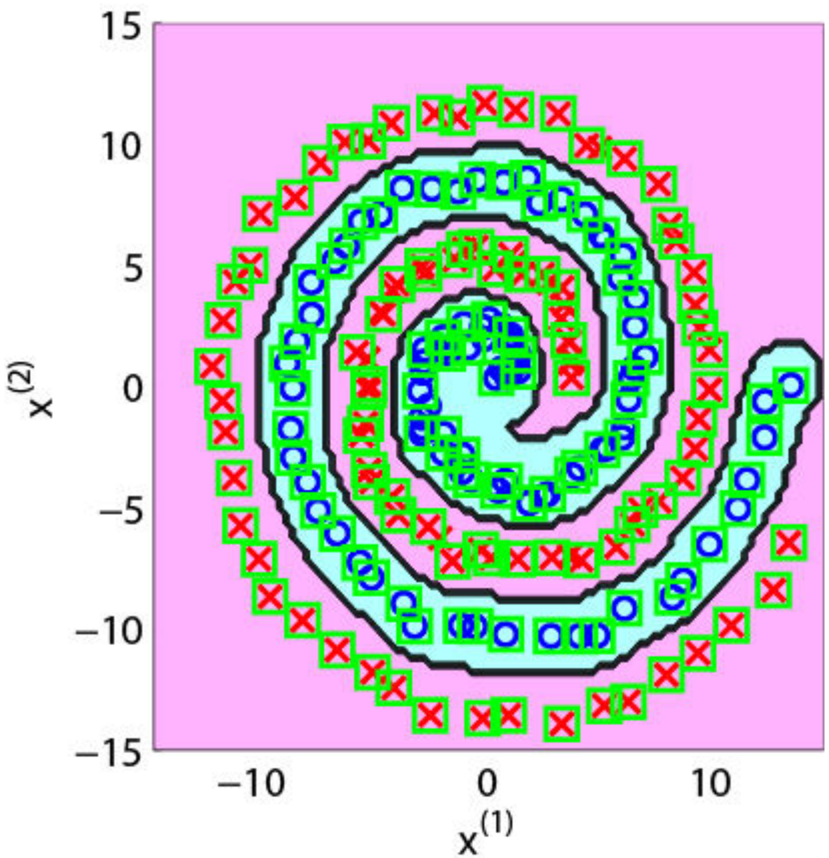
- ヒンジ損失とℓ_2-制約の和を最小にする

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^b}\left[\sum_{i=1}^{n}\max\left(0, 1-f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)y_i\right)+\frac{\lambda}{2}\boldsymbol{\theta}^\top\boldsymbol{R}\boldsymbol{\theta}\right]$$

- 特に以下の場合をサポートベクトル分類器とよぶ
  - $f_{\boldsymbol{\theta}}(\boldsymbol{x})=\sum_{j=1}^{n}\theta_j K(\boldsymbol{x},\boldsymbol{x}_j)+\theta_0$
  - $\boldsymbol{R}=\boldsymbol{K}$ $\qquad K_{i,j}=K(\boldsymbol{x}_i,\boldsymbol{x}_j)$

- 優れたソフトウェアが多数公開されている:
  http://www.support-vector-machines.org/

実行例



- 複雑なデータもきちんと分離できている

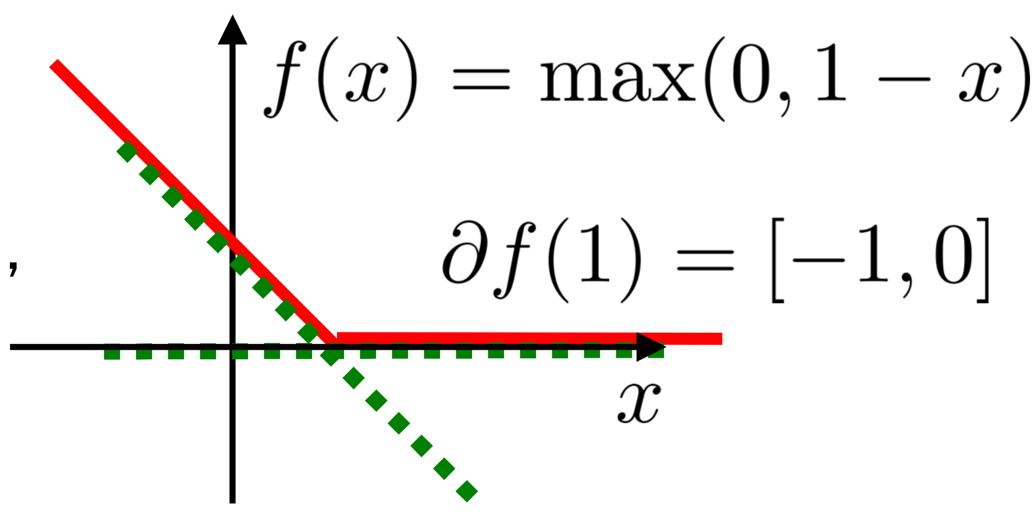
劣勾配法

- 凸関数 f の $\boldsymbol{x}'$ での劣勾配(sub-gradient)とは，全ての $\boldsymbol{x} \in \mathbb{R}^d$ に対して次式を満たす $\boldsymbol{\xi}$：

$$f(\boldsymbol{x}) \geq f(\boldsymbol{x}') + \boldsymbol{\xi}^\top (\boldsymbol{x} - \boldsymbol{x}')$$

  - f が微分可能なとき，$\boldsymbol{\xi} = \nabla f(\boldsymbol{x}')$
  - 上式を満たす $\boldsymbol{\xi}$ 全体を $\partial f(\boldsymbol{x}')$ で表し劣微分(sub-differential)とよぶ

- 劣勾配法：
  - 勾配法において，微分不可能な点では，劣微分のどれかの値を用いる



数学演習

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^b}\left[\sum_{i=1}^{n}\max\left(0,1-f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)y_i\right)+\frac{\lambda}{2}\boldsymbol{\theta}^\top\boldsymbol{R}\boldsymbol{\theta}\right]$$

$$f_{\boldsymbol{\theta}}(\boldsymbol{x})=\sum_{j=1}^{n}\theta_j\exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{x}_j\|^2}{2h^2}\right)$$

- ヒンジ損失の $\boldsymbol{\theta}$ に関する劣微分

$$\partial\max\left(0,1-f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)y_i\right)$$

を求めよ

$$\partial\max\left(0,1-z\right)=\begin{cases}-1 & z<1\\ [-1,0] & z=1\\ 0 & z>1\end{cases}$$

解答例

$$t_j = \partial_j \max\left(0, 1 - f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)y_i\right) \quad f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) = \sum_{j=1}^{n} \theta_j \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}_j\|^2}{2h^2}\right)$$

- $1 - f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)y_i > 0$ のとき $t_j = -y_i \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{2h^2}\right)$
- $1 - f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)y_i = 0$ かつ $y_i = +1$ のとき

$$t_j = \left[-\exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{2h^2}\right), 0\right]$$

- $1 - f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)y_i = 0$ かつ $y_i = -1$ のとき

$$t_j = \left[0, \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{2h^2}\right)\right]$$

- $1 - f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)y_i < 0$ のとき $t_j = 0$

劣勾配アルゴリズム

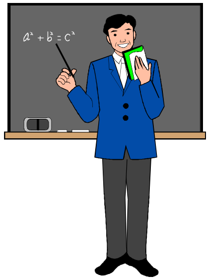
$$\boldsymbol{\theta} \longleftarrow \boldsymbol{\theta} - \varepsilon \left(\sum_{i=1}^{n} \partial \max\Big(0, 1 - f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) y_i\Big) + \lambda \boldsymbol{R}\boldsymbol{\theta} \right)$$

$\varepsilon > 0$:ステップ幅

■ 劣勾配:

$$1 - f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) y_i > 0 \quad \Longrightarrow \quad \partial_j \max\Big(0, 1 - f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) y_i\Big) = -y_i \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{2h^2} \right)$$

$$1 - f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) y_i \leq 0 \quad \Longrightarrow \quad \partial_j \max\Big(0, 1 - f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) y_i\Big) = 0$$



講義の流れ

1. ヒンジ損失（８章）
2. サポートベクトル分類の元々の導出（８章）
3. サポートベクトル分類の解の性質（８章）

サポートベクトル分類器の正統派の導出

- 超平面分類器
- VC理論
- マージン最大化
- ソフトマージン
- カーネルトリック

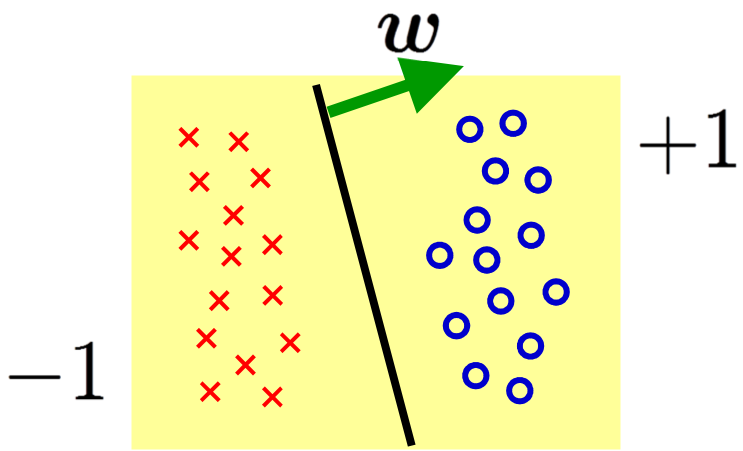
超平面分類器

- 標本空間を超平面で分離する.

$$f_{\boldsymbol{w},b}(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + b$$

$$\boldsymbol{w} = (w^{(1)}, \ldots, w^{(d)})^\top$$

$$\boldsymbol{x} = (x^{(1)}, \ldots, x^{(d)})^\top$$

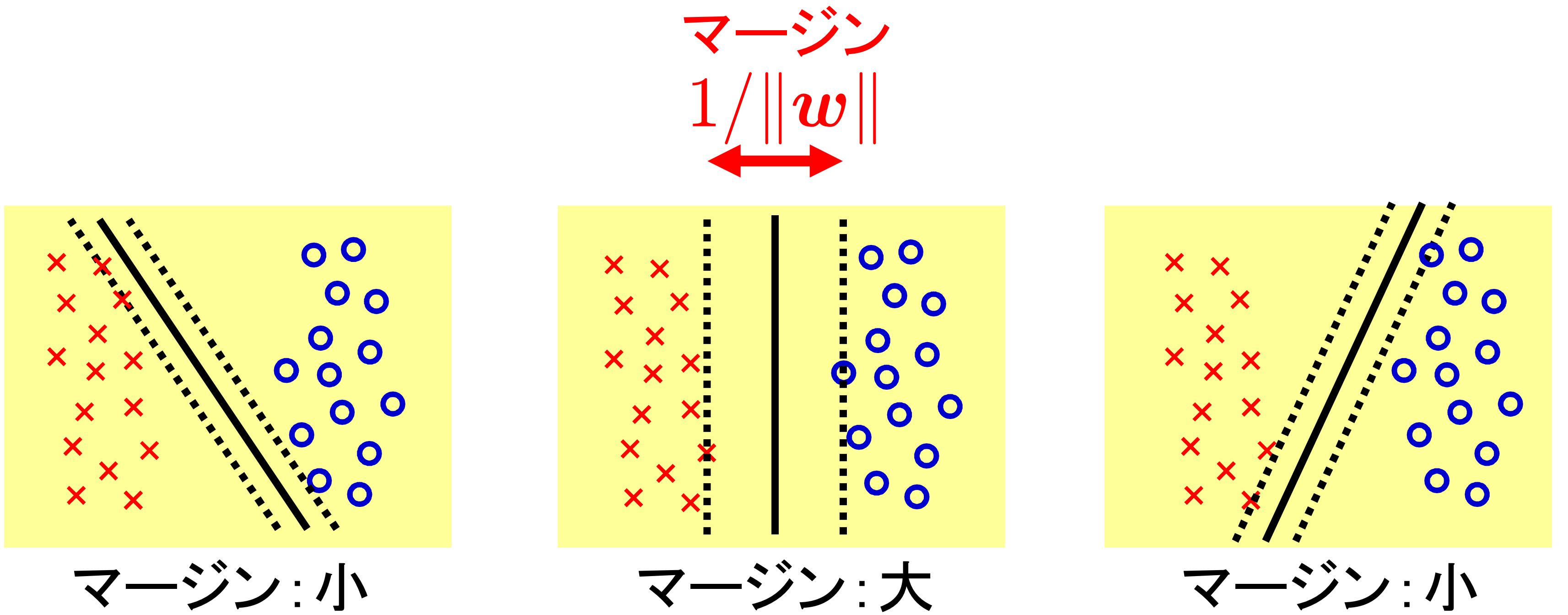


$$\text{find } \boldsymbol{w}, b$$

$$\text{such that} \quad y_i f_{\boldsymbol{w},b}(\boldsymbol{x}_i) \geq 1 \quad \text{for } i = 1, \ldots, n.$$

マージン

■ マージン : 二つのクラスの“隙間”の大きさ



$$f_{\boldsymbol{w},b}(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + b$$

$$y_i f_{\boldsymbol{w},b}(\boldsymbol{x}_i) \geq 1 \quad \text{for } i = 1, \ldots, n.$$

Vapnik-Chervonenkis理論

- 汎化誤差：$R[\widehat{f}] = \iint I(\widehat{f}(\boldsymbol{x}) \neq y) p(\boldsymbol{x}, y) d\boldsymbol{x} dy$
- 経験誤差：$R_{\mathrm{emp}}[\widehat{f}] = \frac{1}{n} \sum_{i=1}^{n} I(\widehat{f}(\boldsymbol{x}_i) \neq y_i)$

$$I(a \neq b) = \begin{cases} 0 & (a = b) \\ 1 & (a \neq b) \end{cases}$$

- 汎化誤差の確率的上界（“VCバウンド”）

$$R[\widehat{f}] \leq R_{\mathrm{emp}}[\widehat{f}] + \sqrt{\frac{1}{n} \left(V \left(\log \frac{2n}{V} + 1 \right) + \log \frac{4}{\delta} \right)}$$

with probability $1 - \delta$

V ：VC次元（分類器の複雑さ）

Vapnik-Chevonenkis理論(続き)

- VCバウンド:

$$R[\widehat{f}] \leq R_{\mathrm{emp}}[\widehat{f}] + \underbrace{\sqrt{\frac{1}{n}\left(V\left(\log\frac{2n}{V}+1\right)+\log\frac{4}{\delta}\right)}}_{\text{VC次元 } V \ (V<n) \text{ の減少に対して単調減少}}$$

- 標本が線形分離可能なとき，経験誤差はゼロ:

$$R_{\mathrm{emp}}[\widehat{f}] = 0$$

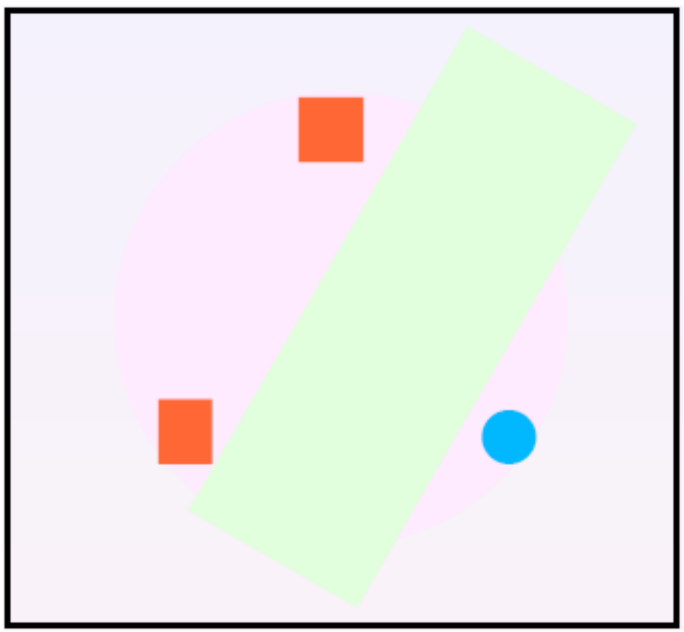
- マージンが大きいほど，VC次元は小さい．

⇒ マージンが最大の超平面識別器が，VC理論においては最適

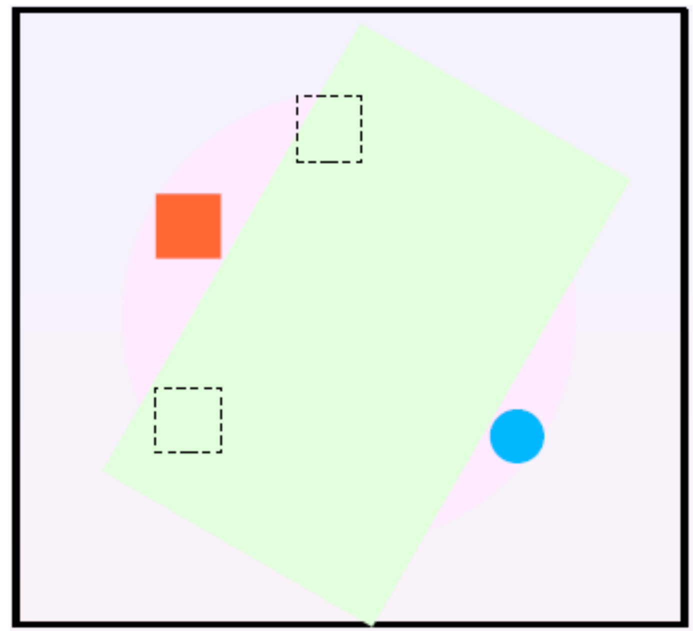
マージンとVC次元

- 一定ノルム内の点の数が多い場合，大きなマージンをもつ超平面識別器が存在できない

マージン小



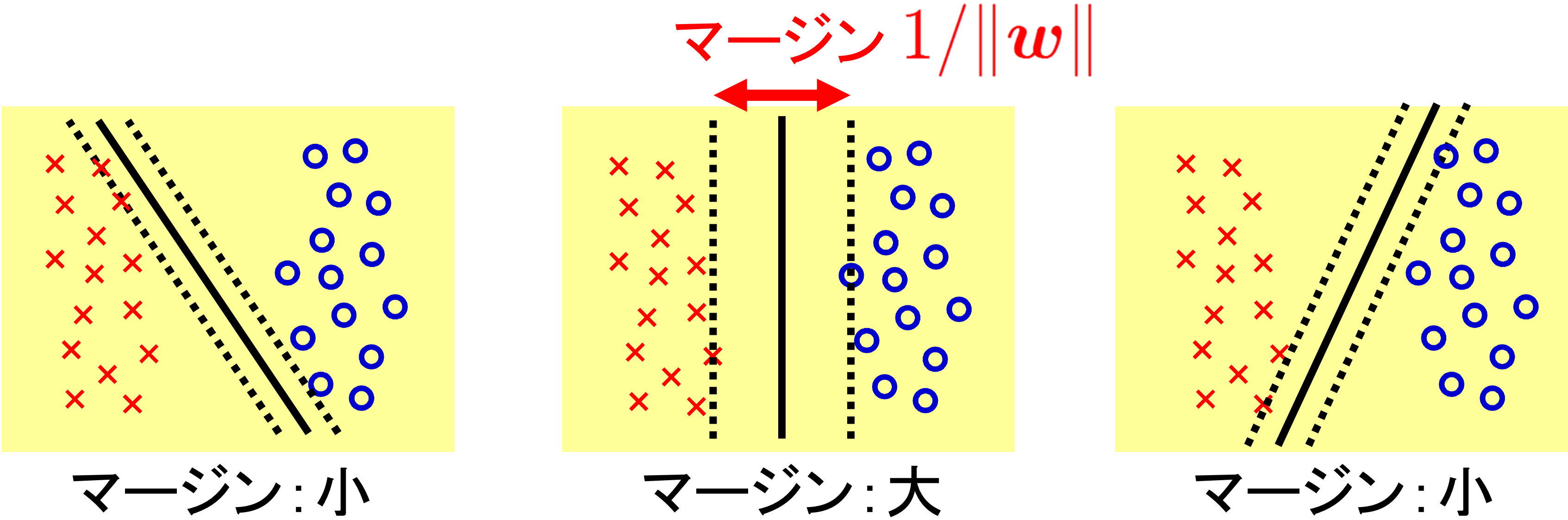
マージン大



[http://www.ide.titech.ac.jp/~yamasita/yylab/06nhk.pdf]

最適超平面分類器

- マージンが最大になるように二つのクラスを超平面で分ける.



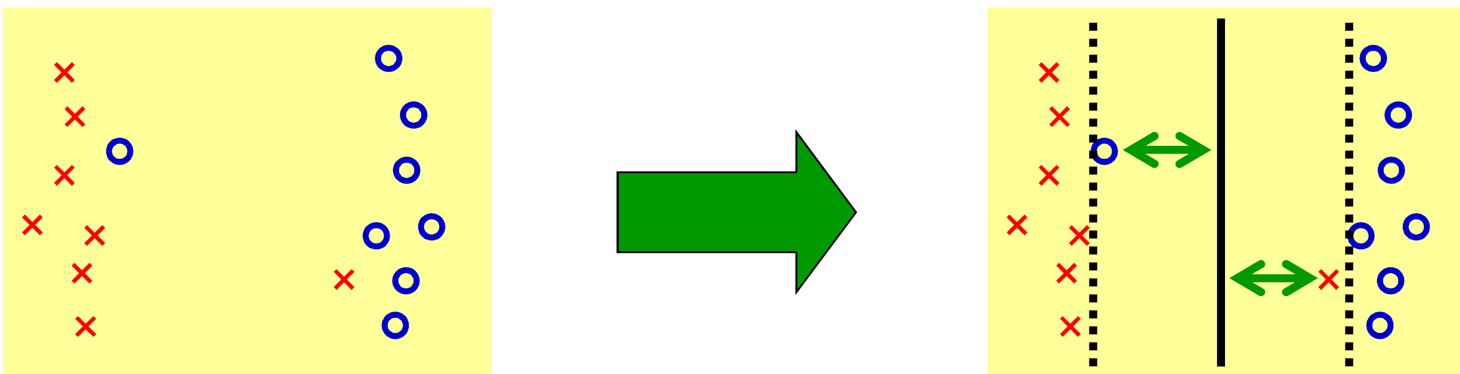
$$\min_{\boldsymbol{w},b} \|\boldsymbol{w}\|^2$$

$$f_{\boldsymbol{w},b}(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + b$$

$$\text{subject to} \quad y_i f_{\boldsymbol{w},b}(\boldsymbol{x}_i) \geq 1 \quad \text{for } i = 1, \ldots, n.$$

ソフトマージン

- 標本が線形分離可能でないときはマージンが定義できない.
- 少しの誤差 ξ_i を許す.



$$\min_{\boldsymbol{w},b,\boldsymbol{\xi}} \|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n}\xi_i$$

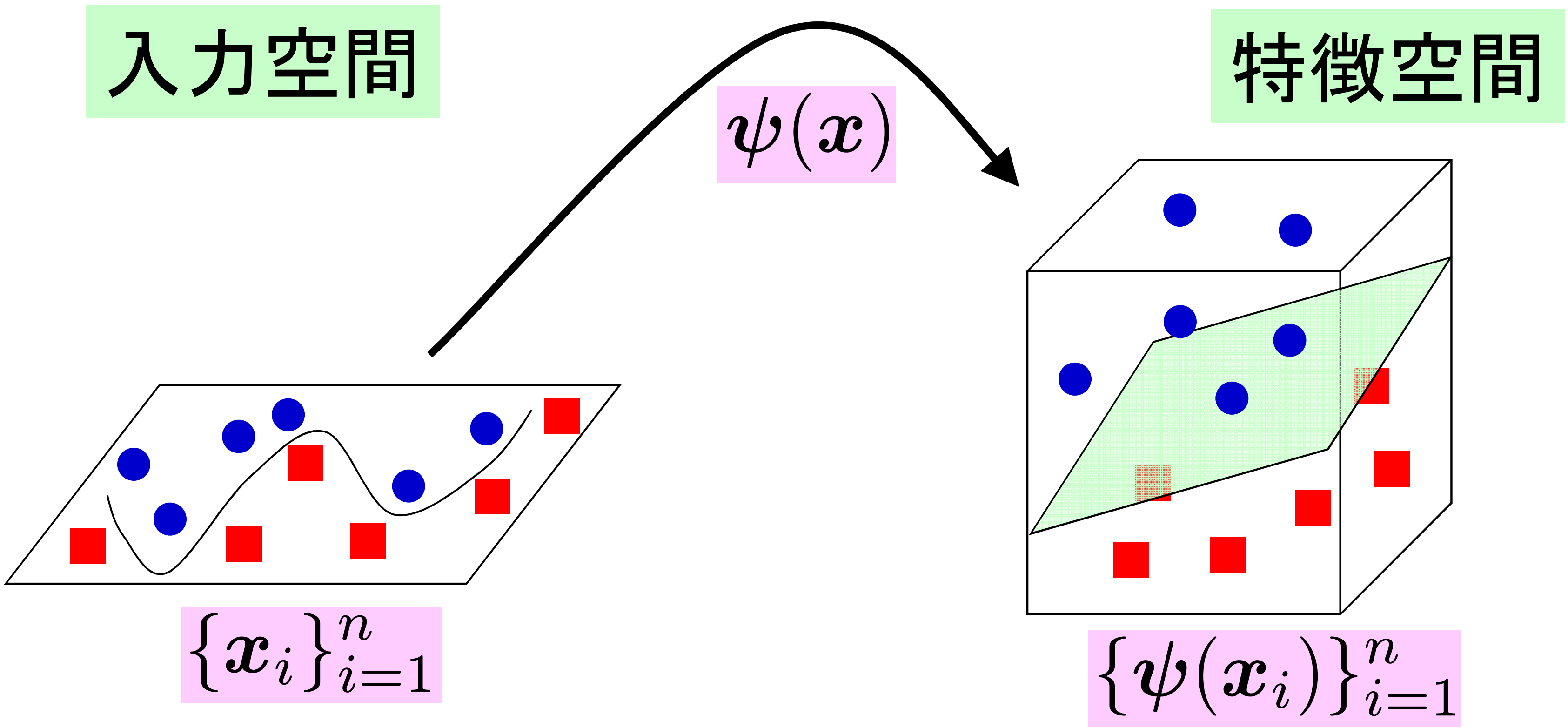
$$f_{\boldsymbol{w},b}(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + b$$

$$\text{subject to} \quad y_i f_{\boldsymbol{w},b}(\boldsymbol{x}_i) \geq 1 - \xi_i$$

$$\xi_i \geq 0 \quad \text{for } i = 1, \ldots, n.$$

非線形への拡張

- 非線形関数 $\psi(\boldsymbol{x})$ で標本を特徴空間へ写像し，特徴空間内でマージン最大の超平面を求める．



カーネルトリック

- 特徴空間内での内積をカーネル関数で計算:

$$\psi(\boldsymbol{x}_i)^\top \psi(\boldsymbol{x}_j) = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

$$\forall \boldsymbol{x}, \boldsymbol{x}', \quad K(\boldsymbol{x}, \boldsymbol{x}') \geq 0$$

例えばガウシアンカーネル

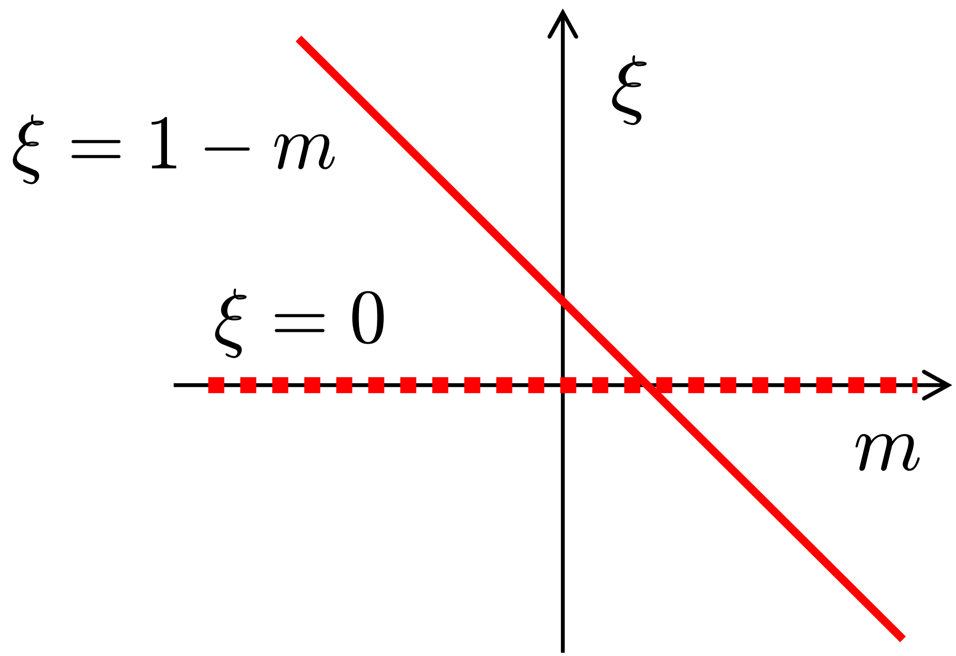
$$K(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\|\boldsymbol{x} - \boldsymbol{x}'\|^2 / (2h^2)\right)$$

- 内積のみで表される線形アルゴリズムは, そのまま非線形に拡張できる.
  - 例: サポートベクトルマシン, 主成分分析, 線形判別分析, K平均クラスタリングなど

ヒンジ損失との関係

- ヒンジ損失は以下のように変形できる

$$\max\{0, 1-m\} = \min_{\xi} \xi \ \ \text{subject to } \xi \geq 1-m,\ \xi \geq 0$$



ヒンジ損失との関係

$$\max\{0, 1-m\} = \min_{\xi} \xi \quad \text{subject to } \xi \geq 1-m, \ \xi \geq 0$$

- ξ_i を消去すると制約なし最適化問題に変形できる:

$$\min_{\boldsymbol{w}, b, \boldsymbol{\xi}} \|\boldsymbol{w}\|^2 + C \sum_{i=1}^{n} \xi_i$$

$$\text{subject to} \quad y_i f_{\boldsymbol{w},b}(\boldsymbol{x}_i) \geq 1 - \xi_i$$

$$\xi_i \geq 0 \quad \text{for } i = 1, \ldots, n$$

$$\Rightarrow \quad \min_{\boldsymbol{w}, b} \|\boldsymbol{w}\|^2 + C \sum_{i=1}^{n} \max\Big(0, 1 - y_i f_{\boldsymbol{w},b}(\boldsymbol{x}_i)\Big)$$

ヒンジ損失との関係

$$\min_{\boldsymbol{w},b} \|\boldsymbol{w}\|^2 + C \sum_{i=1}^{n} \max\Big(0, 1 - y_i f_{\boldsymbol{w},b}(\boldsymbol{x}_i)\Big)$$

$$f_{\boldsymbol{w},b}(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{\psi}(\boldsymbol{x}) + b$$

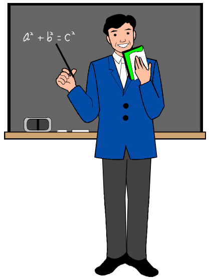
- $\boldsymbol{w} = \sum_{j=1}^{n} \theta_j \boldsymbol{\psi}(\boldsymbol{x}_j)$ とおけばヒンジ損失最小化と等価

$$\min_{\boldsymbol{\theta}} \left[\sum_{i=1}^{n} \max\Big(0, 1 - f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) y_i\Big) + \lambda \boldsymbol{\theta}^\top \boldsymbol{K} \boldsymbol{\theta} \right]$$

$$C = 1/\lambda$$

$$\boldsymbol{\psi}(\boldsymbol{x}_i)^\top \boldsymbol{\psi}(\boldsymbol{x}_j) = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{j=1}^{n} \theta_j K(\boldsymbol{x}, \boldsymbol{x}_j) + \gamma$$



講義の流れ

1. ヒンジ損失（８章）
2. サポートベクトル分類の元々の導出（８章）
3. サポートベクトル分類の解の性質（８章）

最適化問題の双対

- 主問題(primal problem):

$$\min_{\boldsymbol{x}\in\mathcal{X}} f(\boldsymbol{x}) \quad \text{subject to } \boldsymbol{g}(\boldsymbol{x}) = \mathbf{0},\ \boldsymbol{h}(\boldsymbol{x}) \leq \mathbf{0}$$

$$\boldsymbol{g}(\boldsymbol{x}) = (g_1(\boldsymbol{x}), \ldots, g_m(\boldsymbol{x}))^\top$$

$$\boldsymbol{h}(\boldsymbol{x}) = (h_1(\boldsymbol{x}), \ldots, h_n(\boldsymbol{x}))^\top$$

- 双対問題(dual problem):
  - 主問題の最適値の下界を最大化する問題
  - 制約条件から目的関数を作る
  - 目的関数から制約条件を作る

ラグランジュ双対

$$f^* = \min_{\boldsymbol{x} \in \mathcal{X}} f(\boldsymbol{x}) \quad \text{subject to } \boldsymbol{g}(\boldsymbol{x}) = \boldsymbol{0},\ \boldsymbol{h}(\boldsymbol{x}) \le \boldsymbol{0}$$

■ラグランジュ関数(Lagrangian):

$$L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = f(\boldsymbol{x}) + \boldsymbol{\lambda}^\top \boldsymbol{g}(\boldsymbol{x}) + \boldsymbol{\mu}^\top \boldsymbol{h}(\boldsymbol{x})$$

■ラグランジュ双対問題(Lagrange dual problem):

$$f^* = \max_{\boldsymbol{\lambda}, \boldsymbol{\mu}} \inf_{\boldsymbol{x} \in \mathcal{X}} L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu})$$

$$\text{subject to } \boldsymbol{\mu} \ge \boldsymbol{0}$$

- 凸最適化問題に対しては最適解が一致
- 双対問題の方が制約が単純

数学演習

$$\min_{\boldsymbol{w},\boldsymbol{\xi}} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n}\xi_i$$

$$f_{\boldsymbol{w}}(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x}$$

（単純化のため
+bを省略）

$$\text{subject to} \quad y_i f_{\boldsymbol{w}}(\boldsymbol{x}_i) \geq 1 - \xi_i$$

$$\xi_i \geq 0 \quad \text{for } i = 1, \ldots, n$$

- サポートベクトルマシンのラグランジュ関数 $L(\boldsymbol{w}, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ を求めよ
  - $\min_{\boldsymbol{x}\in\mathcal{X}} f(\boldsymbol{x})$ subject to $\boldsymbol{g}(\boldsymbol{x}) = \boldsymbol{0}, \ \boldsymbol{h}(\boldsymbol{x}) \leq \boldsymbol{0}$

    のラグランジュ関数は

$$L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = f(\boldsymbol{x}) + \boldsymbol{\lambda}^\top \boldsymbol{g}(\boldsymbol{x}) + \boldsymbol{\mu}^\top \boldsymbol{h}(\boldsymbol{x})$$

解答例

$$L(\boldsymbol{w}, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n}\xi_i$$

$$-\sum_{i=1}^{n}\alpha_i\left(y_i\boldsymbol{w}^\top\boldsymbol{x}_i - 1 + \xi_i\right) - \sum_{i=1}^{n}\beta_i\xi_i$$

数学演習

■サポートベクトルマシンの双対最適化問題

$$\max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \inf_{\boldsymbol{w},\boldsymbol{\xi}} L(\boldsymbol{w},\boldsymbol{\xi},\boldsymbol{\alpha},\boldsymbol{\beta})$$

$$\text{subject to } \boldsymbol{\alpha} \geq \mathbf{0} \text{ and } \boldsymbol{\beta} \geq \mathbf{0}$$

から $\boldsymbol{w},\boldsymbol{\xi},\boldsymbol{\beta}$ を消去せよ

■ヒント: $\frac{\partial L}{\partial \boldsymbol{w}} = \mathbf{0}, \quad \frac{\partial L}{\partial \boldsymbol{\xi}} = \mathbf{0}$

$$L(\boldsymbol{w},\boldsymbol{\xi},\boldsymbol{\alpha},\boldsymbol{\beta}) = \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n}\xi_i$$

$$-\sum_{i=1}^{n}\alpha_i\left(y_i\boldsymbol{w}^\top\boldsymbol{x}_i - 1 + \xi_i\right) - \sum_{i=1}^{n}\beta_i\xi_i$$

解答例

- $$\frac{\partial L}{\partial \boldsymbol{w}} = 0 \quad \Longrightarrow \quad \boldsymbol{w} = \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x}_i$$
- $$\frac{\partial L}{\partial \xi_i} = 0 \quad \Longrightarrow \quad \alpha_i + \beta_i = C, \ \forall i = 1, \ldots, n$$
- これらより

$$\widehat{\boldsymbol{\alpha}} = \mathop{\mathrm{argmax}}_{\boldsymbol{\alpha}} \left[\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^\top \boldsymbol{x}_j \right]$$

$$\text{subject to } 0 \leq \alpha_i \leq C \text{ for } i = 1, \ldots, n$$

双対最適化問題

- 双対問題は二次計画(quadratic program; QP):

$$\widehat{\boldsymbol{\alpha}} = \underset{\boldsymbol{\alpha}}{\mathrm{argmax}} \left[\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^\top \boldsymbol{x}_j \right]$$

$$\text{subject to } 0 \leq \alpha_i \leq C \text{ for } i = 1, \ldots, n$$

- 主問題の解:

$$\widehat{\boldsymbol{w}} = \sum_{i=1}^{n} \widehat{\alpha}_i y_i \boldsymbol{x}_i$$

$$\frac{\partial L}{\partial \boldsymbol{w}} = 0 \implies \boldsymbol{w} = \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x}_i$$

最適性条件

$$\min_{\boldsymbol{x}\in\mathcal{X}} f(\boldsymbol{x}) \quad \text{subject to } \boldsymbol{g}(\boldsymbol{x}) = \boldsymbol{0},\ \boldsymbol{h}(\boldsymbol{x}) \leq \boldsymbol{0}$$

$$\boldsymbol{g}(\boldsymbol{x}) = (g_1(\boldsymbol{x}), \ldots, g_m(\boldsymbol{x}))^\top$$

$$\boldsymbol{h}(\boldsymbol{x}) = (h_1(\boldsymbol{x}), \ldots, h_n(\boldsymbol{x}))^\top$$

- 最適解の必要条件（凸の場合は必要十分条件）：
  - $\nabla f(\boldsymbol{x}^*) + \boldsymbol{\lambda}^{*\top}\nabla \boldsymbol{g}(\boldsymbol{x}^*) + \boldsymbol{\mu}^{*\top}\nabla \boldsymbol{h}(\boldsymbol{x}^*) = \boldsymbol{0}$
  - $\boldsymbol{g}(\boldsymbol{x}^*) = \boldsymbol{0}$
  - $\boldsymbol{h}(\boldsymbol{x}^*) \leq \boldsymbol{0}$
  - $\boldsymbol{\mu}^* \geq \boldsymbol{0}$
  - $\mu_i^* h_i(\boldsymbol{x}^*) = 0, \quad i = 1, \ldots, n$ ←相補性条件

カルーシュ・キューン・タッカー
(KKT)条件
(Karush-Kuhn-Tucker conditions)

相補性条件

$$\min_{\boldsymbol{w},\boldsymbol{\xi}} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n}\xi_i \qquad f_{\boldsymbol{w}}(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x}$$

$$\text{subject to} \quad y_i f_{\boldsymbol{w}}(\boldsymbol{x}_i) \geq 1 - \xi_i$$

$$\xi_i \geq 0 \quad \text{for } i = 1, \ldots, n$$

- サポートベクトルマシンの双対最適化問題の相補性条件は

$$\alpha_i(y_i \boldsymbol{w}^\top \boldsymbol{x}_i - 1 + \xi_i) = 0 \qquad \beta_i \xi_i = 0$$

$$\text{for } i = 1, \ldots, n$$

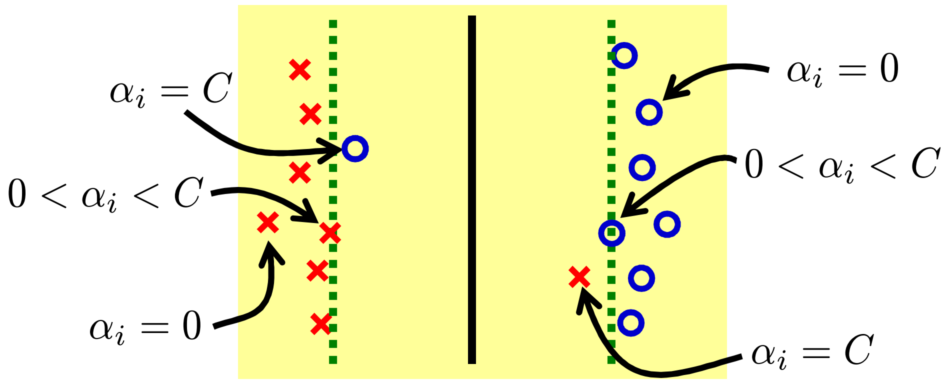
相補性条件（続き）

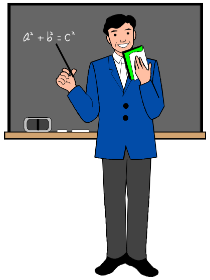
$$\alpha_i(y_i \boldsymbol{w}^\top \boldsymbol{x}_i - 1 + \xi_i) = 0 \qquad \beta_i \xi_i = 0$$

- 相補性条件より，以下の性質が成り立つ：

1. $\alpha_i = 0 \implies y_i \boldsymbol{w}^\top \boldsymbol{x}_i \geq 1$
2. $0 < \alpha_i < C \implies y_i \boldsymbol{w}^\top \boldsymbol{x}_i = 1$
3. $\alpha_i = C \implies y_i \boldsymbol{w}^\top \boldsymbol{x}_i \leq 1$
4. $y_i \boldsymbol{w}^\top \boldsymbol{x}_i > 1 \implies \alpha_i = 0$
5. $y_i \boldsymbol{w}^\top \boldsymbol{x}_i < 1 \implies \alpha_i = C$

導出は宿題





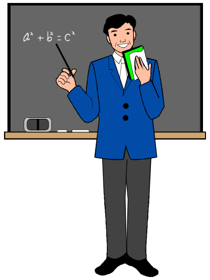
講義の流れ

1. ヒンジ損失（８章）
2. サポートベクトル分類の元々の導出（８章）
3. サポートベクトル分類の解の性質（８章）

まとめ

- 0/1-損失の代理としてはヒンジ損失が適切
- サポートベクトル分類は，元々はマージン最大化原理に基づいた学習法として導出
- 分類問題：
  - 入力に関する線形モデル＋最小二乗学習
    ＝フィッシャーの線形判別分析
  - カーネルモデル＋正則化ヒンジ損失最小化
    ＝サポートベクトルマシン

→ 良い方法には様々な解釈がある



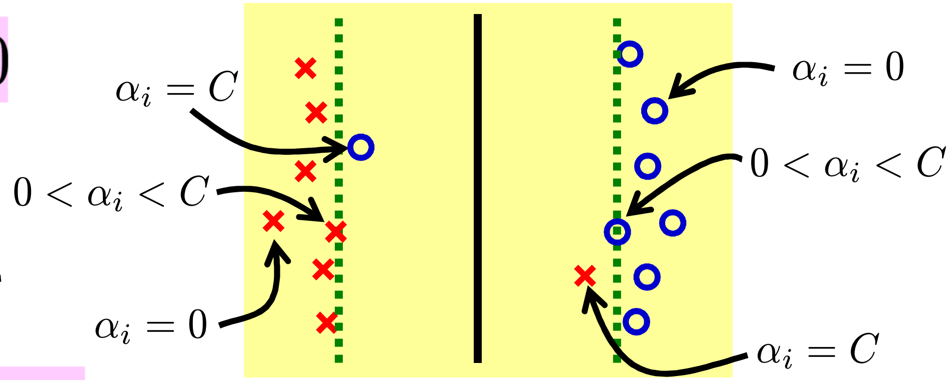
次回の予告

- 確率的分類（１０章）
- 系列データの分類（１１章）

宿題1

相補性条件

- $\alpha_i(y_i \boldsymbol{w}^\top \boldsymbol{x}_i - 1 + \xi_i) = 0$
- $\beta_i \xi_i = 0$



より，以下の性質を示せ

1. $\alpha_i = 0 \implies y_i \boldsymbol{w}^\top \boldsymbol{x}_i \geq 1$
2. $0 < \alpha_i < C \implies y_i \boldsymbol{w}^\top \boldsymbol{x}_i = 1$
3. $\alpha_i = C \implies y_i \boldsymbol{w}^\top \boldsymbol{x}_i \leq 1$
4. $y_i \boldsymbol{w}^\top \boldsymbol{x}_i > 1 \implies \alpha_i = 0$
5. $y_i \boldsymbol{w}^\top \boldsymbol{x}_i < 1 \implies \alpha_i = C$

ヒント: 以下の条件を利用する

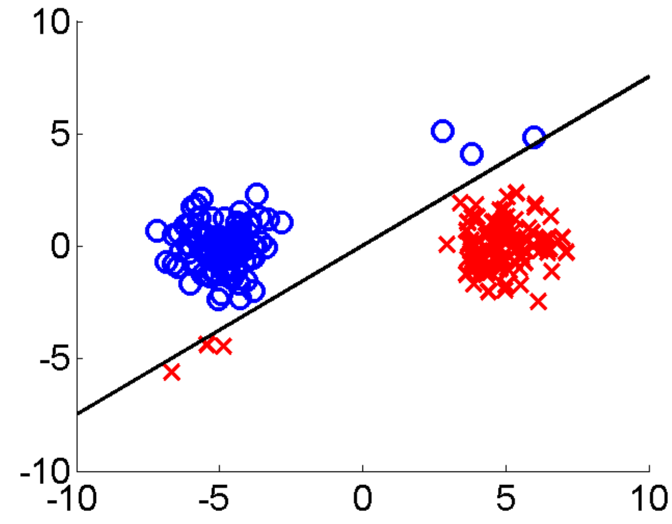
$$y_i \boldsymbol{w}^\top \boldsymbol{x}_i - 1 + \xi_i \geq 0 \qquad \xi_i \geq 0 \qquad \alpha_i + \beta_i = C$$

宿題2

- 線形モデル

$$f_{\boldsymbol{w},b}(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + b$$

に対するサポートベクトルマシンの劣勾配アルゴリズムを実装せよ




```
clear all; rand('state',0); randn('state',0);
n=200; x=[randn(1,n/2)-5 randn(1,n/2)+5; randn(1,n)]';
x(:,3)=1;
y=[ones(n/2,1);-ones(n/2,1)]; y(1:3)=-1; y(n/2+1:n/2+3,1)=1;
x(1:3,2)=x(1:3,2)-5; x(n/2+1:n/2+3,2)=x(n/2+1:n/2+3,2)+5;

<<< From x and y, learn 3-dimensional vector w >>>

figure(1); clf; hold on; axis([-10 10 -10 10]);
plot(x(y==1,1),x(y==1,2),'bo');
plot(x(y==-1,1),x(y==-1,2),'rx');
plot([-10 10],-(w(3)+[-10 10]*w(1))/w(2),'k-');
```